

## AP STAT: CHAPTER 3

## CATEGORICAL DATA

**\*\*MAKE A PICTURE!\*\***

**First, create a frequency table**

Example: number of students at CB South in each grade:

Grade	TOTAL
10	534
11	552
12	515

Proportion = decimal      10<sup>th</sup>:  $534 / \text{total} = 0.30$   
Percent = %      30%

Frequency = # count

Relative frequency = %

Distribution (of a variable)-

values of variable & how often  
we see each value

Examples:

bar chart  
pie chart

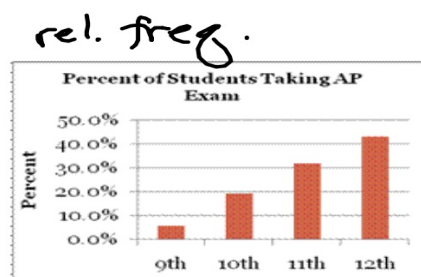
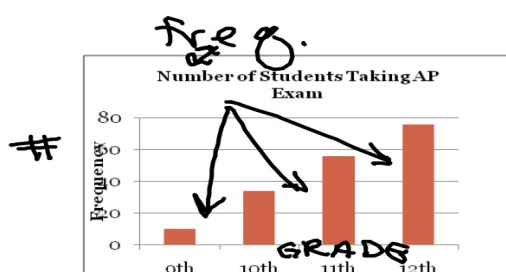
stemplot  
histogram

boxplot  
dotplot

### Categorical Distributions:

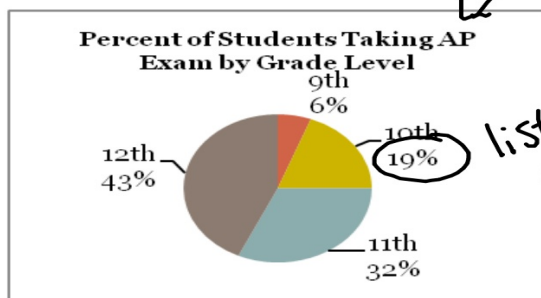
#### 1. Bar Chart

label  
axes



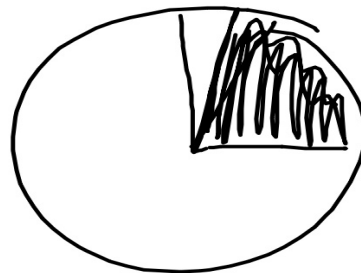
Pie charts

% label



listed as %

7/20



### 3. Contingency tables (aka Two-Way tables)

	Fr.	So	Jr.	Sr.	total
M	cell	cell	.		
F	cell	----			
total					n

$$n = 100$$

Identify:

- Row variable *Gender*
- Column variable *Grade*
- Values of the variable *M/F*
- Total *(n)*
- # of Cells *8*
- Totals *Fr. / So / Jr / Sr*

**Example:** Hospitals

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
	2100	800

- What percent of people died?
- Of those people that went to Hospital A, what percent died?
- Of those people who went to Hospital B, what percent died?
- Of those people who died, what percent went to Hospital A?
- What percent of people died and went to Hospital B?

$$\frac{79}{2900} = 2.72\%$$

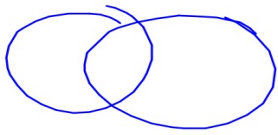
$$\frac{63}{2100} = 3\%$$

$$\frac{16}{800} = 2\%$$

$$\frac{63}{79} = 79.75\%$$

overlap

$$\frac{16}{2900} = 0.55\%$$



2 types of Distributions for Categorical Variables — 2 way tables

1) MARGINAL DISTRIBUTIONS

- How to make:

margins  $\div$  n (total)

- Looking for ...

overall % in each category

- ALWAYS ...

in %

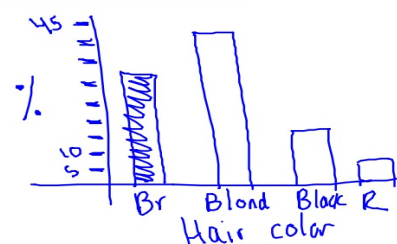
value

- Example: Hair color vs. Gender

	Brown	Blonde	Black	Red	Total
MALE	26	24	10	3	63
FEMALE	20	35	12	6	73
TOTALs	46	59	22	9	136

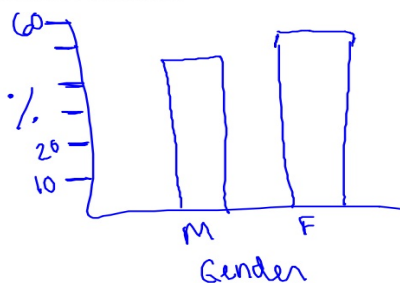
- Find the marginal distribution for the HAIR COLOR variable

$$\begin{aligned} \text{Brown} &= 46/136 = 33.82\% \\ \text{Blonde} &= 59/136 = 44.38\% \\ \text{Black} &= 22/136 = 16.18\% \\ \text{Red} &= 9/136 = 6.62\% \end{aligned}$$



- Find the marginal distribution for the GENDER variable

$$\begin{aligned} \text{Male} &= 63/136 = 46.32\% \\ \text{Female} &= 73/136 = 53.68\% \end{aligned}$$



- Represented Visually: BAR CHART

1) **CONDITIONAL DISTRIBUTIONS**

- Look at ... one variable

Then look at ... each value individually

- Break down ... each value into its pieces
- ALWAYS ... %



- Example: Hair Color vs. Gender

Hair color

- Find the conditional Distribution for the HAIR COLOR variable

	Brown	Blonde	Black	Red	Total
MALE	26	24	10	3	63
FEMALE	20	35	12	6	73
TOTALs	46	59	22	9	136

Brown  
M:  $26/46 = 56.52\%$   
F:  $20/46 = 43.48\%$

Blonde  
M:  $24/59$   
F:  $35/59$

Black  
M:  $10/22$   
F:  $12/22$

Red  
M:  $3/9 =$   
F:  $6/9 =$

- Find the conditional Distribution for the GENDER variable

Male

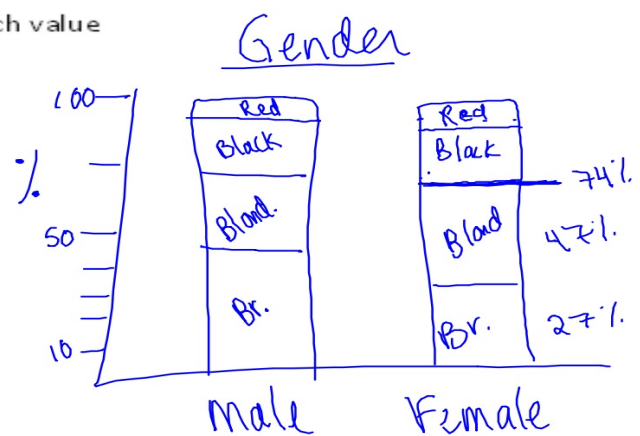
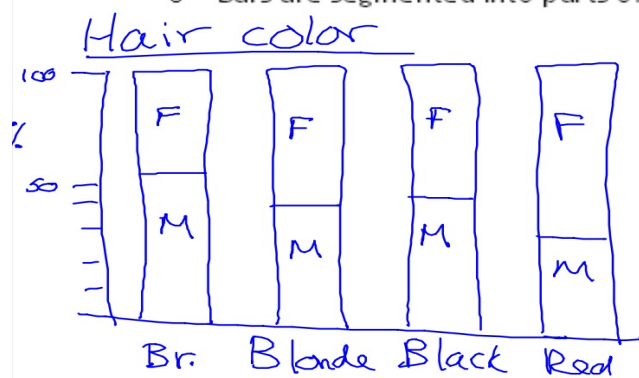
Br.  $26/63 = 41.27\%$   
Blon  $24/63 = 38.1\%$   
Bla  $10/63 = 15.87\%$   
Red  $3/63 = 4.76\%$

Female

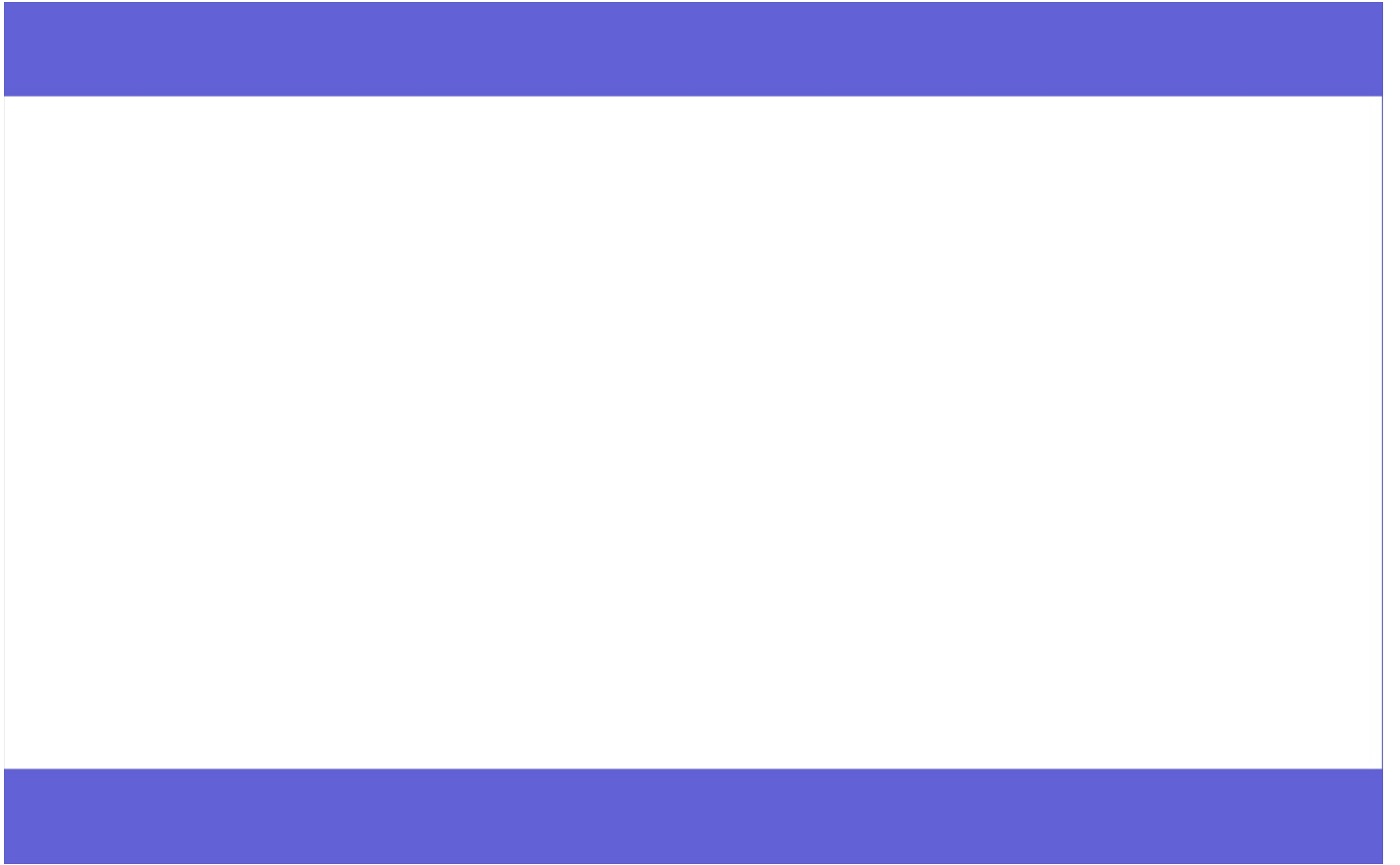
Br  $20/73 = 27.4\%$   
Blond.  $35/73 = 47.95\%$   
Bla  $12/73 = 16.44\%$   
Red  $6/73 = 8.22\%$

- Represented visually: SEGMENTED (or STACKED) BAR GRAPH

- Each bar = 100%
- Values of variable on the x-axis
- Bars are segmented into parts of each value



Try worksheet 3A on your own! Work  
with someone else in the class if you  
want



Warm Up: worksheet 3B, only do letters (a) and (d)

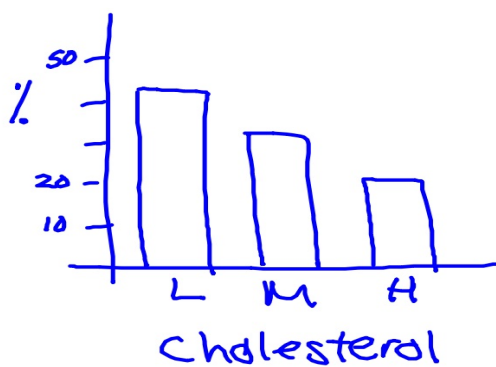
	Low chol.	Med. chol.	high chol.
NE heart attacks	29	17	18
Fatal heart attack	19	20	9

Ⓐ marginal for cholesterol level & bar chart

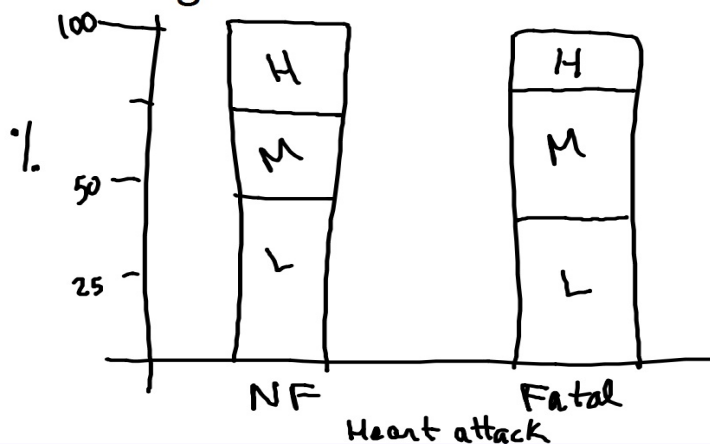
Ⓑ conditional for heart attack & bar chart

### Answers to worksheet 3B:

(a) low: 42.9%  
med: 33.0%  
high: 24.1%

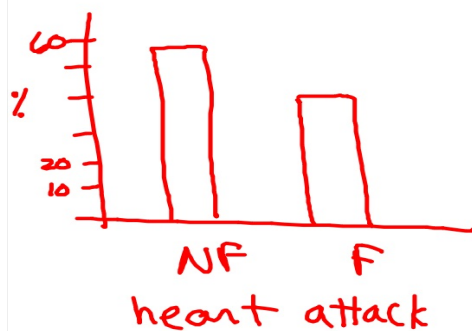


(d)            NE                      Fatal  
Low: 45.3%                      39.6%  
Med: 26.6%                      41.7%  
High: 28.1%                      18.8%



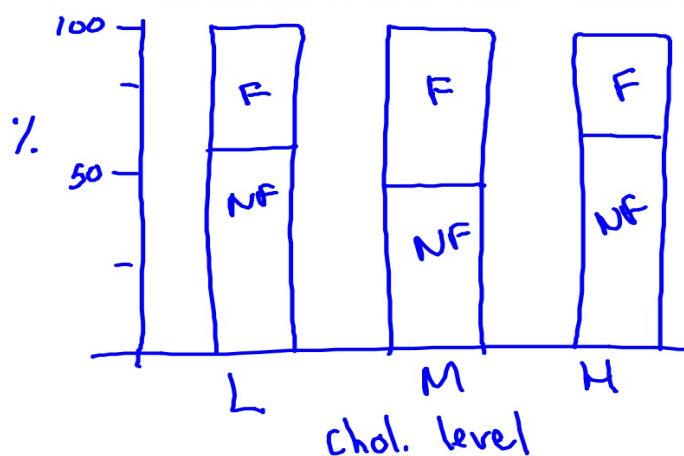
# Worksheet 3B continued:

(b) NF: 57.1%  
F: 42.9%



(c)

	low	med	high
NF	60.4%	45.9%	66.7%
F	39.6%	54.1%	33.3%



### Worksheet 3C:

A study was made to compare year in high school with preference for vanilla or chocolate ice cream with the following results.

	Vanilla	Chocolate	
Freshman	20	10	30
Sophomore	24	12	36
Junior	18	9	27
Senior	22	11	33
	84	42	126

(a) Marginal for grade level:

Fr: 23.8%

So: 28.57%

Jr: 21.43%

Sr: 26.19%

(b) Marginal for flavor:

Vanilla: 66.7%

Chocolate: 33.33%

(c) Conditional for grade level:

FROSH

V: 66.67%

C: 33.33%

SOPH

V: 66.67%

C: 33.33%

JR

V: 66.67%

C: 33.33%

SENIOR

V: 66.67%

C: 33.33%



(d) Conditional for flavor

Vanilla

Fr: 23.81%  
So: 28.57%  
Jr: 21.43%  
Sr: 26.19%

Chocolate

Fr: 23.81%  
So: 28.57%  
Jr: 21.43%  
Sr: 26.19%

independent =  
no association

marginal for one var  
= condit. for other var.

Worksheet 3C: question 2

- (a) 16%
- (b) 16%
- (c) NO
- (d) Marginal for Prediction:
  - Pred. stay: 75%
  - Pred. leave: 25%

- (g) Conditional for Actual:

**actual stay:**

Pred. stay: 75%  
Pred. leave: 25%

**actual leave:**

pred. stay: 75%  
pred. leave: 25%

✓  
Indep.

(e) Marginal for actual:

Actual stay: 84%  
Actual leave: 16%

(f) conditional for predicted:

***Pred. stay***

act. stay: 84%  
act. leave: 16%

***Pred. leave***

act. stay: 84%  
act. leave: 16%

	Brown	Blonde	Black	Red	Total
MALE	26	24	10	3	63
FEMALE	20	35	12	6	73
TOTALS	46	59	22	9	136

Is Gender indep.  
of hair color?

overall % males  $\approx$  46%

% males red: 33%  


---

 overall % brown: 34%

***Worksheet 3D: Simpson's paradox***

1) Kulp: 80%      Cardi: 90%

3) GOOD CONDITION

Kulp: 98.3%      Cardi: 96.7%

4) POOR CONDITION

Kulp: 52.5%      Cardi: 30%